

Attachment 13

Final Summary Report

External Letter Peer Review of FDA's

Quantitative Consumer Research on Cigarette Health Warnings Required by the Family Smoking Prevention and Tobacco Control Act

November 19, 2019

Contract No. HHSF223201700015B
BPA No. 10

Prepared for:

Stephanie L. Redus, M.S. and David B. Portnoy, Ph.D., MPH
U.S. Food and Drug Administration
Center for Tobacco Products
10903 New Hampshire Ave.
Silver Spring, MD 20993

Peer Reviewers:

Joseph N. Cappella, Ph.D.
Jennifer C. Morgan, Ph.D.
John P. Pierce, Ph.D.
Kurt M. Ribisl, Ph.D.
William G. Shadel, Ph.D.
James F. Thrasher, Ph.D.

Prepared by:



Versar, Inc.
6850 Versar Center
Springfield, VA 22151

II. RESPONSE TO CHARGE QUESTIONS

1. Is the document logical and clear?

The document is written in clear English and for the most part the presentation is logical and concise. The problem is that it is not clear why there are eight different aims addressed in this study. It is well into the document before the reader learns that there are eight aims and, for many of them, the relevance to the study purpose is far from clear. It needs a theoretical framework from which the relevance of each aim is presented.

2. Does the executive summary accurately reflect the content of the overall document?

Similar to the above statement, the executive summary does present how the findings of the study relate to the purpose of the study. This is written in clear and understandable English. However, this means that the numerous additional aims that are addressed in this study are not mentioned at all. The authors can't have it both ways. Either make the case for including the additional aims or delete the analyses relating to them from the report.

3. Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?

This reader was very impressed with the presentation of the overall study design, the choice of sample and experimental methods used for this study. As above, the problem was the additional aims that were not well justified and, indeed, the study did not have the power to address all of them. This puts at least one of these aims into the 'exploratory aim' category. While it may meet some internal needs of the FDA to obtain this information at the same time as conducting this study, it is not clear why the final aim should be included in this report.

4. Is the methodology used appropriate to address the study's purpose?

The study is well designed and is appropriate to address the study's purpose.

5. Are the stimuli used appropriate given the study's purpose?

The experimental design is excellent and allows the assessment of the study aims related to the study purpose. The rationale for presentation of study stimuli is well presented.

6. Are the outcomes measured appropriate given the study's purpose?

No. Only a subset of the aims are appropriate for the study's purpose, as it is currently framed. This reviewer suggests that the wording of the study purpose be qualified so that the new understandings relate to a potential change in smoking behavior. The study lacks a theoretical framework section that demonstrates why each of the outcomes measured is relevant to the study's purpose.

7. Are the study participants included appropriate given the study's purpose?

Yes, the choice of the study population and efforts taken to recruit them are appropriate. This is not a representative sample of the population, but it doesn't need to be. The goal is to test how a diverse population respond to study messages and the allocation between study groups is unbiased.

8. *Is the analytic approach appropriate given the design and purpose of the study?*

This reviewer was impressed with the study analytic approach which was dictated by a very good experimental design.

9. *Are results presented consistent with the analytic approach?*

Yes, the results to all the study aims are thoughtfully and clearly presented.

10. *Are there any concerns with the results presented?*

Yes, some of the aims appear to be unrelated to the study's purpose as currently written. The report tries to make the case for this as it presents the results. However, the danger is that this looks a little like *post-hoc rationalization*. There was no theoretical framework section outlining the relevance of each of these additional measures, so the reader has no knowledge of whether the hypothesis for each aim is met or not. Nor what a negative finding might mean to the overall purpose of the study. Indeed, it is not until the second study that the point of some of the aims becomes clear (some change with graphic warning labels but not with text warning labels). A combined theoretical framework is very much needed for these studies.

11. *Are potential limitations of the study appropriately identified?*

Yes, the limitations section in this report is appropriate.

12. *Are the conclusions drawn from the study well supported by the data presented?*

Yes, the conclusions outlined in the executive summary are well supported by the analyses and clearly relate to the study's purpose as it is currently stated. However, there is a lack of consideration of the meaning of the additional aims that are not directly related to the current (limited) specification of the study purpose. See issues above on the need for a detailed theoretical framework for the study.

III. *Specific Observations on Experimental Study on Warning Statements for Cigarette Graphic Health Warnings: Study 1 Report*

None provided.

Quantitative Consumer Research on Cigarette Health Warnings: Study 2

I. GENERAL IMPRESSIONS

There is a lot of overlap in the methods section for Study 2 with that of Study 1. I suggest that there be a re-positioning of these two studies so that you can combine the two methodology sections into a single presentation.

The two studies could also be combined as they are somewhat hierarchical. The first study investigates the response to new and improved text messages. The second study investigates how graphic warning labels can enhance the response to these text messages. A very important conclusion from these two studies is that text messaging alone does not achieve changes in health beliefs, but, when combined with graphic imaging, health beliefs are influenced. It is not possible to do this when the material is presented as two separate, and apparently independent, studies.

This report needs a much better section on the theory of health communications that incorporates why changing health beliefs is important and a step above the communication-persuasion achieved with the text only communication. I suggest a version of McGuire's communication-persuasion matrix that incorporates emotive processing in the behavior change model. The linear nature of this model is too simplistic, however, there is a hierarchy that is important.

For this application, exposure is controlled by the excellent experimental design that enables identification of associations with the different processes of communication-persuasion. First of all, the participant needs to identify that the message has new information and that it is understandable. When this new, understandable information is considered factual, it can lead to new cognitions (thinking about risks). However, it is important to behavior change that this new knowledge leads to a modification of health beliefs that are retrievable at the time of performance of the addictive behavior. When the exposure generates an emotive response associated with the risks of use, then it is likely that there will be greater change in health beliefs. The goal of putting graphic warning labels on cigarette packaging is that the image will continue to generate an emotive response which will be a cue to retrieve these health beliefs every time the person reaches for a cigarette.

Both studies address how messages can assist people to make changes to their behavior and I would incorporate them into two sections of the same report.

II. RESPONSE TO CHARGE QUESTIONS

1. *Is the document logical and clear?*

To this reviewer the weakness in the report is the lack of an explicit theory that addresses why the different measures are undertaken and what each construct is expected to achieve in terms of the final outcome. Is this behavior change, or is it just the first step towards this – the identification of new knowledge?

2. *Does the executive summary accurately reflect the content of the overall document?*

The executive summary presents the findings clearly, however, it lacks the theoretical model that will help the reader understand the importance of the findings to the overall goal. Is it just new knowledge or new knowledge that will assist the smoker in the behavior change process? This reviewer thinks that the second study is built on the first study and that they should be presented as one. The second study shows how graphic images lead to more advanced processing than text only messages that results in a change in health beliefs. Thus, these warnings will be more associated with increased probability of behavior change. At the present, there is no attempt to address why graphic warning labels are a step-above the textual warnings, although the findings can directly address this.

3. *Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?*

The study design is enlightened and allows for a rigorous testing of the multiple applications of warning labels. The stimuli are well chosen. The sample is an appropriate size. Allocation to the respective groups is by a minimization method first introduced in the biostatistical literature in the 1970s and when grouped with quotas for key sampling units ensures that there is comparability across study groups with appropriate representation of the key population components. This allows this research to be undertaken using the online panels. The downside is that it is not a randomization procedure, however, the assumptions of the statistical testing are robust enough that this procedure does not introduce significant bias. The analyses are appropriate and the results meaningful.

4. *Is the methodology used appropriate to address the study's purpose?*

Yes. The methodology demonstrates careful consideration of the design principles adapted for use with online panels. The only thing that is lacking is a justification for the choice of the two-week follow-up. This is an easy addition as this time-point is a trade-off between the need to go beyond short-term memory recall while keeping the timing close-enough to minimize loss to follow-up.

5. *Are the stimuli used appropriate given the study's purpose?*

Yes. Each graphic image was appropriately matched with a text message used in the first study. In the instances in which there was not an appropriate match, the investigators used random assignment. This is an optimal approach.

6. *Are the outcomes measured appropriate given the study's purpose?*

The problem with the outcome measures is that there is no presentation of a theory on why they are important. Of particular concern is the relevance of health beliefs. This is the key difference between the text only vs. graphic warning labels. See above.

External Letter Peer Review of Quantitative Consumer Research on Cigarette Health Warnings Required by the Family Smoking Prevention and Tobacco Control Act

Reviewer #2

Quantitative Consumer Research on Cigarette Health Warnings: Study 1

I. GENERAL IMPRESSIONS

My primary comments are indicated below, but a few stand out: (1) The need for some overarching conceptual framework to bring coherence to the outcomes assessed and interpretation of results; (2) stronger justification for the measures used, including information on the validity of measures, especially novel ones; (3) consideration of prior research to determine meaningful effect sizes and power; (4) stronger justification for the phase 2 belief assessment; (5) stronger justification for the analytic approaches and inclusion of sensitivity analyses to evaluate the consistency of results under different specifications (e.g., approach to randomly selecting comparative TCA warning).

II. RESPONSE TO CHARGE QUESTIONS

1. *Is the document logical and clear?*

Yes. However, to make it even clearer, I recommend that the authors integrate a figure to illustrate the protocol steps, including when stimuli were shown and the timing of specific measures.

I recommend eliminating redundancy between the methodology report and the results report, which includes a LOT of the same information.

2. *Does the executive summary accurately reflect the content of the overall document?*

Clarify that the stimuli are just text (i.e., no imagery/pictures). Otherwise, it is a bit confusing since the title of the report indicates “graphic” warnings.

Add a justification for the three primary outcomes selected, including an indication of the constructs that they are supposed to measure. I would also mention the “secondary” outcomes since they are generally important constructs for evaluating message effects in a brief experiment like this (see comments in the outcome section for concerns about designating some measures as secondary without any theoretical or empirical justification).

For phase 2, clarify this statement so that the reader understands what it means without going to the methods section: “Participants assigned to the treatment conditions viewed one of several different combinations of 9 revised warning statements.”

Not clear what this means: “After viewing the 9 warning statements, all participants answered questions about their beliefs about the link between smoking and each of the health consequences presented in the warning statements.” Did everyone answer the same questions, some of which included health effects that were on the warnings that they

evaluated and some of which were not? Or did people just evaluate outcomes that were on the warnings they evaluated. What was done has implications for the analysis and its interpretation (see below).

For phase 2, did the control group get the same health belief questions asked as in phase 1? Clarify.

Include descriptive information for the 4 of 15 revised statements that were higher than the standard warnings on thinking about risks given its importance in predicting cessation.

Clarify what is being compared in the health beliefs summary at the end of the results section.

3. *Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?*

The information provided is mostly sufficient. I would add the following material to enhance the clarity of the information presented:

- Integrate a figure to illustrate the protocol steps, including when stimuli were shown and when specific measures were used.
- Include a figure with the actual stimuli as shown to participants.
- Provide more information on how revised warning statements were selected
- Include more information about measurement (see comments in the next section)

4. *Is the methodology used appropriate to address the study's purpose?*

In section 2.1, experimental design: The rationale for the approach used for the treatment conditions should be spelled out, as it is not clear. In particular, justify showing only one revised statement and eight of the original warning statements.

Provide a justification for evaluating beliefs in two separate moments with different measurement approaches.

In the methods section or the limitations, the authors could do a better job of citing literature indicating the consistency of results from online studies of warning responses and those that use either physical packs in brief experiments or that compare responses to warnings in online studies with those smokers have after policies are rolled out (e.g., (1) Hammond D, Thrasher JF, Reid JL, Driezen P, Boudreau C, Arillo-Santillán E. Perceived effectiveness of pictorial health warnings among Mexican youth and adults: A population-level intervention with potential to reduce tobacco-related inequities. *Cancer Causes and Control*. 23 (Suppl): 57-67. 2012; (2) Huang L, Thrasher JF, Reid J, Hammond D. Predictive and external validity of a pre-market study to determine the most effective pictorial health warning label content for cigarette packages. *Nicotine & Tobacco Research*. 18(5):1376-1381. 2016; (3) Thrasher JF, Carpenter M, Andrews JO, Gray KM, Alberg AJ, Navarro A, Friedman DB, Cummings KM. Cigarette warning label policy alternatives and smoking-related health disparities. *American Journal of Preventive Medicine*. 43(6):590–600. 2012; (4) Hammond D, Reid JL, Driezen P, Boudreau C. Pictorial health warnings on cigarette packs in the United States: an experimental evaluation of the proposed FDA warnings. *Nicotine Tob Res*. 2013

Jan;15(1):93-102).

Power calculations that adjust for false discovery rates (Benjamini & Hochberg, 1995) are important given the great number of comparisons made. However, the authors would ideally provide citations and empirical justification for the anticipated effect size (difference of 0.5 and standard deviation of 1) given the substantial body of research in this area. Otherwise, it is hard to determine if the study is over or underpowered. This information will also be useful when considering the unanticipated equal allocation of sample to treatment and control groups, especially given the significantly lower power found for the equal allocation scenario relative to the optimized allocation with a larger control group. The authors may be able to address this concern by using the literature to show the effect size, including meta-analyses.

5. *Are the stimuli used appropriate given the study's purpose?*

Yes. The stimuli seem pretty standard for online studies of warnings; however, it would be clearer if the report showed example stimuli to illustrate what the stimuli looked like to participants (which they do not currently do).

Topics for the new warnings generally capture outcomes about which there is likely to be lower awareness in the general population. How these topics were selected should be clarified, as there is no information about this issue in the report.

6. *Are the outcomes measured appropriate given the study's purpose?*

The conceptual model(s) that orients this study is underdetermined and never clearly defined. At the end of a single sentence, the authors cite a bunch of studies to support their measurement strategy. The report would be stronger if it provided citations separately for each measure used and an indication of how the construct it measures fits within a framework for message effects and/or the conceptualization of “understanding” (given the FDA mandate to increase public understanding). The primary outcomes of “new information” and “self-reported learning” have some face validity as potential indicators of understanding as knowledge accumulation. Still, the report would ideally cite studies with more convincing validity for these indicators. The other primary outcome, “thinking about risks,” has substantial predictive validity and relevant studies should be cited (e.g., many studies of adult smokers have shown that this response to warnings is associated with downstream cessation attempts). Some researchers consider this measure as indicative of message engagement or elaboration, which is a more standard term in communication research.

The authors should provide some justification for selecting some indicators as primary and others as “secondary,” ideally based on the conceptual model that orients the study. For example, “informativeness” and “factuality” appear to overlap with the conceptualization of understanding. Why are they secondary? Why is credibility not primary, especially given that the messages are mostly about less well-known smoking-related outcomes? Decisions to treat these as secondary appears even more arbitrary after reading Study 2, where all measures are treated equally (i.e., no distinctions are made between primary and secondary measures).

Phase 1 health belief questions use 5-point Likert type response options, which, in my opinion, does not really “fit” with the idea of a belief. In section 4.3, the report states: “Conceptually, the response categories for a Likert response scale represent an underlying belief continuum”. The report would ideally provide some justification for this approach to measuring beliefs (as opposed to attitudes, frequencies or other constructs for which a continuum makes sense conceptually and is more standard).

The authors should include citations for each specific measure, as most are not standard (e.g., learning, new knowledge, informativeness, factuality).

For phase 1, it is not clear to me if all participants answered the same belief questions or, as was implied in the executive summary, that this list included only the beliefs associated with the warnings that they evaluated. If the latter, it is not clear how alphas were calculated (due to incomplete data). If the former, provide an explanation for why overall beliefs were evaluated.

Better justify how asking health beliefs in phase 2 ads meaningful information.

Phase 2 health belief questions appears to be a series of 22 questions with check boxes. Better justify the creation of summative measures across all knowledge outcomes if we are interested in determining sensitivity to content that is included in warnings to which they are exposed.

The belief questions may be more about memory and test taking skills than “understanding”.

The authors should consider this as a potential limitation, especially since they show the stimuli multiple times and evaluate beliefs at two distinct moments.

7. *Are the study participants included appropriate given the study’s purpose?*

Yes: The focus on adolescents who smoke and are susceptible to smoke is standard for this kind of study, as is the inclusion of established adult smokers. However, it is not clear why ever-smokers who are susceptible were excluded (and only never smokers susceptible were included). Quotas for young adult smokers and older adult smokers is also appropriate given differential effects of warnings found for these populations and concerns about trying to influence young adults before they become too addicted.

8. *Is the analytic approach appropriate given the design and purpose of the study?*

See above comments on measurement of outcomes that are relevant to analyses.

The authors could create quasi-control groups with the treatment groups that are not shown a revised warning with the health outcome of interest. This would increase power for phase 1 measures, which would be particularly beneficial for evaluating statistical significance within the 3 key subgroups.

As mentioned above, the authors would ideally provide citations and empirical justification for the anticipated effect size given the substantial body of research in this area. Contrasts like the one that I mention in the prior comment would allow for greater power.

Table 4-2 shows alpha for a number of different subscales of beliefs. It is unusual to use alpha for assessing the internal consistency for two questions. The methods literature with which I am familiar indicates the need for a minimum of three questions for alpha to be meaningful.

Throughout the section on hypothesis statements (starting at 4-11), the hypotheses are presented as directional (e.g., "...treatment condition > ...control condition"), but the statistical tests are indicated to be "two-sided". Either make the hypotheses NOT EQUAL or change the language around the statistical tests to indicate "one-sided".

4-13 is where the analysis approach that involves randomly selecting a TCA statement to serve as the control for the revised statements without clear parallel content. This results in random selection of all responses to just one TCA statement for each comparison. Given the very limited pool of TCA statements, this approach to random selection risks integrating a systematic bias around consumer responses to the particular statement that is selected. For example, the TCA statement on "addiction" addresses a concept that is notoriously difficult to communicate and is often evaluated as less effective than well-known disease outcomes. Nevertheless, it was randomly selected as the comparison for the statement about "macular degeneration". A more robust comparison would involve a random selection from all TSA statements or even the grand mean of responses to all TSA statements, which would help iron out systematic idiosyncrasies around the specific topic that is randomly selected from a pool of only 9 possible TSA statements. This could be done when comparing treatment and control, as well as within person comparisons.

4.3.3, Phase 1, part 2: Hypotheses and Analyses - The hypothesis phrase "average or level health belief score" is not clear, particularly the meaning of the term "level".

The analyses that involve evaluating whether the score was significantly higher than "not at all" (or 0 on a 0 to 7 scale) for learning, believability, and informativeness is unorthodox and should be better justified. I do not see this analysis as adding anything meaningful to what is already done with means and linear regression.

Given the study design, provide a clearer justification for creating and evaluating a summative measure of smoking-related health consequences, as well as an overall health consequences measure. This justification should speak to issues around how we would expect beliefs to be higher for participants who are exposed to warning statements that address that specific belief (compared to participants who are not exposed to statements with that content). The summative measures used seem to reflect a broad conceptualization of risk perception (that goes beyond the content of the warnings) and should be justified. The more specific domains around secondhand smoke and pregnancy consequences do a better job of mapping onto specific warning content and therefore do not raise this issue.

Appendix A indicates that a "control" belief was evaluated for each of the three domains of health consequences. To control for social desirability and acquiescence biases, it would be standard to include this as an adjustment variable in analyses that involve these beliefs. Was that done?

[The Prior comments are from the methods report. Unless otherwise indicated, what follows

is on the results report, although issues I raise above are pertinent to the background and methods sections of the results.]

9. *Are results presented consistent with the analytic approach?*

Yes, the results are presented in a way that is consistent with the analytic approach.

Tables use a $p < 0.05$ for unadjusted results. This is consistent with a one-tailed test, not a two-tailed test. As I mention my comment about the analytic approach, the wording of the hypothesis is directional but stating that two-tailed tests were used suggests non-directional hypotheses.

3.6 Phase 2 results - The description mentions respondents who “saw only revised statements,” but even these saw TCA statements in phase 1. After reading the results, I am still not clear how phase 2 results add anything meaningful. This should be clarified.

10. *Are there any concerns with the results presented?*

I am not clear why results are shown for tests that are not adjusted for multiple testing. The inclusion of these more problematic assessments does not seem to add any information of import and weakens the presentation by raising the question “Why have this information?”.

For all tables that involve the comparison with a randomly selected TCA statement (e.g., Table 3-4), include a footnote or indication of the topic of the randomly selected warning in the table (if you do not choose to follow my recommendations above to do a different comparison).

11. *Are potential limitations of the study appropriately identified?*

When raising the issue of the ecological validity of responses to online images of warning statements, the authors could cite studies that indicate convergent validity (see comment above).

Concerns about generalizability could be addressed with studies showing that patterns of response in experiments are generally consistent across population sources. Also, as a sensitivity analysis, the authors could consider weighting observations so that they are more similar to the profile of smokers and nonsmokers in the general population.

The limitations should do a better job of describing potential measurement error, citing the validity (or lack thereof) for the measurement approaches used. This could include considerations of content validity around approaches for measuring “understanding.” More broadly, there may be alternative conceptualizations of “understanding” that would encompass embodied/experiential understanding. This kind of understanding may be stronger for smoking-related diseases associated with sensory perceptions from smoking (e.g., lung, throat, mouth, heart). Some evidence suggests that smokers perceive warning labels for these well-known outcomes as more effective than warnings for less-well known outcomes. Warnings may serve as reminders about this embodied understanding.

12. Are the conclusions drawn from the study well supported by the data presented?

The summary of findings should report on the significance of comparisons after adjustment for multiple tests. For example, I believe that the last sentence of the first paragraph in this section discusses 8 of 16 comparisons as higher for revised TCA statements, when I think it is only four after adjustment.

The summary would ideally discuss patterns of findings across indicators, rather than treating them one at a time, which loses the broader patterns.

The primary concern that I have is around the apparent prioritization of “primary outcomes” on “learning” and “new knowledge”. I am unfamiliar with prior research showing the validity and meaningfulness of the outcomes used to measure understanding (i.e., meeting FDA’s mandate). I am more familiar with indicators like the one used for “thinking about risks,” for which there is substantial evidence of predictive validity for cessation attempts across a variety of warning label policies and sociocultural contexts. Looking at the data for the revised statement on erectile dysfunction, for example, it generates more knowledge but lower thinking about risks and lower believability – which would recommend against its use. There are many other examples of inconsistent results, as well. These concerns about the validity of measures, primary vs secondary outcomes (voiced in the section above on outcomes), and the consistency of patterns across indicators of effect become particularly important when interpreting the results to inform Study 2. The documents would ideally be better linked, so that conclusions from Study 1 clearly inform the selection of stimuli for use in Study 2.

III. Specific Observations on *Experimental Study on Warning Statements for Cigarette Graphic Health Warnings: Study 1 Report*

None provided.

Quantitative Consumer Research on Cigarette Health Warnings: Study 2

I. GENERAL IMPRESSIONS

I have provided my comments below, emphasizing those that are most important. A few stand out: 1. the need for some overarching conceptual framework to bring coherence to the outcomes assessed and interpretation of results; 2. stronger justification for the warnings selected, ideally based on Study 1 results and prior research; 3. stronger justification for the measures used, including information on the validity of measures, especially novel ones; 4. consideration of prior research to determine meaningful effect sizes and power; 5. stronger justification for the analytic approaches (e.g., combining four SG warning control groups for comparison instead of creating more comparable comparison groups) and inclusion of sensitivity analyses to evaluate the consistency of results under different specifications (e.g., population weights; adjustment for variables that account for differential attrition).

II. RESPONSE TO CHARGE QUESTIONS

1. *Is the document logical and clear?*

Yes. However, there is significant redundancy between the methods and results reports that would ideally be deleted.

2. *Does the executive summary accurately reflect the content of the overall document?*

Yes. However, it would be useful to indicate which of the 15 warning statements are revised and which are original statements from the TCA. Also, I would put the actual stimuli in the table that shows the warnings.

Rather than listing the outcomes, the summary would be more compelling if these outcomes were somehow linked to the overarching conceptualization of “understanding” or to established frameworks on message effects. My comments about specific outcome measures from Study 1 apply to this Study 2, as well, since many of the same outcomes are used. There are also some new outcomes assessed in Study 2 that should be justified (e.g., understandability, helpfulness) by linking to this conceptual framework. For the executive summary, this could be done briefly, with more detailed description and justification in the background and measurement sections.

3. *Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?*

The two studies should be better linked, so that results from Study 1 clearly inform the selection of stimuli for use in Study 2. This is not currently done, and there is no justification for the selection of stimuli for Study 2.

4. *Is the methodology used appropriate to address the study’s purpose?*

The methods are generally appropriate.

There should be a stated rationale for the session 2 exposure, which I assume is to simulate naturalistic, repeated exposure to warnings that would happen in standard implementation periods.

Power calculations that adjust for false discovery rates (Benjamini & Hochberg, 1995) are important given the great number of comparisons made. However, the authors would ideally provide citations and empirical justification for the anticipated effect sizes given the substantial body of research comparing pictorial and text only warnings. Otherwise, it is hard to determine if the study is over or underpowered. This information will also be useful when considering the recommendation to conduct additional analyses that compare specific SG warnings with comparable GHWs, as well as power for evaluating subgroup analyses.

5. *Are the stimuli used appropriate given the study's purpose?*

Stimuli seem appropriate, but the justification for selecting warning statements and associated pictorial imagery should be spelled out. Ideally, this would build on Study 1 findings and/or the scientific literature.

When describing the stimuli in the text, the report should describe differences in the size and placement of the control vs. treatment statements.

6. *Are the outcomes measured appropriate given the study's purpose?*

As I commented for Study 1, the conceptual model(s) that orients measurement for Study 2 is never clearly defined. At the end of a single sentence in section 2.3 Instrument Development, the authors cite a bunch of studies to support their measurement strategy. The report would be stronger if they provided citations separately for each measure used and an indication of how it fits under a framework for message effects and/or the conceptualization of “understanding” (given the FDA mandate to increase public understanding). My comments around outcomes that are also used in Study 1 apply to Study 2, so I will not repeat them. However, there are also some new measures for Study 2 that I have not seen before and that also would benefit from some information about their validity and prior use. For example, is “Understandability” the same thing as “clarity”? If so, that construct is relatively common in studies of perceived effectiveness and could be cited as such. The new question on attention grabbing is commonly used and should be cited.

The authors should include citations for each specific measure, as most are not standard (e.g., learning, new knowledge, informativeness, factuality). This will help with interpretation of results.

It is not clear how the measurement of B1 (Before today, had you heard about the specific smoking-related health effect described in the warning?) was done for warnings where multiple health effects are mentioned in the warning statements. Were participants asked to interpret the entire gestalt of the warning statement, including if it mentioned multiple health outcomes (e.g., SG warning on “lung cancer, heart disease, emphysema and may complicate pregnancy”)?) If so, there is a lack of “fit” between some warning statements and question B1. There is also a lack of fit when the warning does not address a specific health effect (i.e., “Quitting smoking now greatly reduces serious risks to your health;” “Cigarette smoke

contains carbon monoxide”). **These issues should be addressed in the limitations.**

Measurement of recall is okay, but the authors would ideally describe the specific type of assessment (which some call “recognition”) and, in the limitation section, potential biases associated with this type of recall vs. other types (e.g., confirmed recall, cued recall). For a good discussion of the strengths and weaknesses of different approaches to self-reported recall, see Niederdeppe J. Conceptual, empirical, and practical issues in developing valid measures of media campaign exposure. *Communication Methods and Measures*, 8(2), 138-161. 2014.

7. *Are the study participants included appropriate given the study’s purpose?*

Generally, yes. The focus on adolescents who smoke and are susceptible to smoke is standard for this kind of study, as is the inclusion of established adult smokers. However, it is not clear why ever-smokers who are susceptible were excluded (and only never smokers susceptible were included). Also, it is not clear why adult nonsmokers were included. The report should justify these decisions.

8. *Is the analytic approach appropriate given the design and purpose of the study?*

My primary concern around the analysis is the decision to combine into a single control group the four distinct groups that were exposed to each of the four different SG warnings. This is a reasonable approach if the concern is to mitigate social desirability and testing effects, but I think it would be stronger to also test differences between specific SG warnings and GHWs that are most comparable. In other words, compare responses to the SG warning on health effects (i.e., ...lung cancer, heart disease, emphysema and may complicate pregnancy) with the GHWs on health effects (perhaps noting which comparisons include the health effects mentioned in both warnings and which do not). The SG warning on quitting can be compared with the GHW on quitting. The SG warning on pregnancy could be compared with the GHW on pregnancy. The SG warning on carbon monoxide is not comparable with any GHWs. At the very least, these more focused analyses could be treated as sensitivity analyses. This focused approach may be most meaningful for the evaluation of measure of “new information” and changes in health beliefs, but I think it would strengthen all analyses.

Power appears reasonable, but raises questions given that the “control” group is actually four distinct groups exposed to four distinct messages. Ideally, power would address comparisons where a specific SG warning is compared with the warnings with similar content (see above comments).

Not sure why results that do not adjust for multiple comparisons are shown. As far as I can tell, they do not provide additional meaningful information (especially as the 95% CIs are shown).

Concerns about generalizability could be addressed somewhat by developing weights for the sample to make it more representative of age/sex/smoking status composition of the general population. This could be used for sensitivity analyses, with consistent results taken as evidence for the likely generalizability.

[My prior comments are in response to the methods report. Unless otherwise indicated, what follows is on the results report, although issues I raise above are pertinent to the background and methods sections of the results.]

9. *Are results presented consistent with the analytic approach?*

The results are consistent with the analytic approach described.

10. *Are there any concerns with the results presented?*

As far as I can tell, there was no assessment of whether the characteristics of the treatment and control groups were significantly different (i.e., whether randomization worked).

The range of days from baseline to completion of session 2 and session 3 should be included. Just the median is currently provided.

Table 3-1 would ideally include statistical tests for differences in sample composition across sessions. While most characteristics appear similar, there appears to be some meaningful attrition by age and smoking status over time.

Differential attrition could be partly addressed by including as control variables the specific age X smoking status variables that appear to differ over time (e.g., a single category for never smoker susceptible adolescents) rather than what I interpret as controlling for age and smoking status as separate adjustment variables.

The assessment of differential attrition by treatment and control groups is only done at the group level. There is no assessment of whether the sociodemographic and smoking status composition of the groups becomes more dissimilar over time. Indeed, such differences could emerge even if the overall attrition rate was the same across groups. The report would be strengthened by this kind of assessment and inclusion of statistical controls for the characteristics that become significantly different over time.

Table 3-3. Show the means for each SG warning done separately, as well as averaged together (see prior comment about analysis of specific SG warnings).

11. *Are potential limitations of the study appropriately identified?*

When raising the issue of the ecological validity of responses to online images of warning statements, the authors could cite studies that indicate convergent validity (see comments from Study 1).

Concerns about generalizability could be addressed by citing studies showing that patterns of response in experiments are generally consistent across population sources. Also, as a sensitivity analysis, the authors could weight observations so that they are more similar to the profile of smokers and nonsmokers in the general population and evaluate the consistency of results.

The limitations should better describe potential measurement issues raised above (e.g., “fit”

of B1 for some warnings; measurement issues described for common measures with Study 1), including implications for interpreting results. See Study 1 comments on considerations of content validity around measurement of “understanding.”

12. Are the conclusions drawn from the study well supported by the data presented?

The lack of an overarching framework for and validity of the outcomes assessed makes it challenging to interpret results, particularly around factualness, which is lower for GHWs than SG warnings. The current explanation is neither based in empirical evidence nor theory.

III. Specific Observations on *Experimental Study of Cigarette Warnings: Study 2 Report*

None provided.

External Letter Peer Review of Quantitative Consumer Research on Cigarette Health Warnings Required by the Family Smoking Prevention and Tobacco Control Act

Reviewer #3

Quantitative Consumer Research on Cigarette Health Warnings: Study 1

I. GENERAL IMPRESSIONS

The comments that I have made in response to the 12 charge questions include various elements that would fall under “general impressions.” I draw them out here in a separate answer.

Both studies are very well done in terms of design and data analysis. The designs selected provide for an appropriate control groups which are the current standards for warnings (Surgeon General in Study 2) or immediate past selections (initial FDA warning labels in Study 1). The data analysis plan is strong and straight forward. It is careful on statistical treatment of the quality of data (e.g. continuous vs. rank order) and especially strong on correcting for experiment-wise error and power considerations. Both documents do a good job in communicating their procedures and the results except as noted regarding Study 1’s complex and difficult design that requires some supplementation with a clear and effective graphical description. If one accepts that the operational measures of learning and novelty are valid measures of their underlying constructs, then there is a clear picture that the new warning texts and warning labels are effective relative to their comparison.

However, I am concerned that the measures deployed – perceived novelty and awareness – are not convincing measures of the underlying constructs that the research is targeting. In Study 1, the researchers do employ a measure of believability and of facticity (opinion vs. fact) finding that the new labels are less believable. In Study 2 the believability measure is not present even though it was diagnostic in Study 1. Both the believability and facticity measures underscore the fact that the new warnings may not be accepted by the target audience. The authors are well aware of this and comment on it in both studies. But coupled with self-report measures of learning and novelty (awareness), the lower levels of acceptance of the labels reduce the overall impact of the results. The implicit rejoinder in the data to the argument that the results are not so convincing is that the warnings affect the acceptance of negative health consequences (i.e. beliefs) in the warning label conditions versus the SG warnings condition and do so over time. But as I note in my comments on Study 2, two concerns arise about these findings. The first is that asking beliefs at baseline before message exposure taints the message processing by focusing respondents’ attention to the messaging in ways that privilege the beliefs being targeted. Second, the beliefs are measured three times reinforcing the warning labels’ content. Third, it is not clear why all beliefs would be affected by a specific warning label as opposed to a more targeted set of outcomes wherein warning label X affects beliefs related to warning label X but not beliefs Y and Z.

In the end, these are both very carefully done studies that adhere closely to the data that has been gathered. This reviewer is raising interpretive considerations that essentially claim that the overall set of results are less convincing than they might be had the same constructs been operationalized differently and slight changes in the design in Study 2 been implemented.

II. RESPONSE TO CHARGE QUESTIONS

1. Is the document logical and clear?

Yes, the document is logical and clear in many respects, most in fact. However, the design of Study 1 is very complicated and unusual, and it took this reviewer multiple readings before it became clear what was taking place and what the exact nature of the protocol was. I would strongly recommend a visual presentation of the protocol to help readers understand what the sequence was and to understand the kinds of questions being asked at various stages along with the warnings to which respondents were exposed at the different stages and phases of the protocol.

2. Does the executive summary accurately reflect the content of the overall document?

Yes, I thought the executive summary was a good representation of the positive findings and procedures discussed in much greater detail in the subsections of the ensuing documents. The only respect in which I thought the executive summary was a little bit misleading -- and this may reflect more of my views about the research than others who read the document differently -- is the presentation of findings about believability. These can be construed as negative and/or problematic for the research and they should be a part of the executive summary. I'm pretty sure that the authors who prepared the work do not agree. They have offered some commentary in Study 1 and definitely in Study 2 about concerns that might be raised about believability and factualness. More about this below for Studies 1 and 2.

3. Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?

Yes, I thought that there was plenty of detailed information about the design, stimuli, sampling methods, and analysis both in the two documents that made up Study 1 plus the supplementary materials. If I was looking for any additional information it would have been about the measures of susceptibility to smoking among adolescents which I couldn't initially find although it does turn up later in the Study 1 and in Study 2. Second, as I mentioned above, I think that the design of the study is complex and difficult to understand and could profit from a careful visual presentation of the protocol. It's not that the information is not present; it is. However, it's just difficult to fully comprehend what. My final point concerns the stimulus materials that are generated for testing. There is some brief discussion of how these materials were generated and while it may not be of value or of interest to understand the sifting and winnowing process here, **this reviewer was a bit perplexed about the sources and topics that generated the new warnings.**

4. Is the methodology used appropriate to address the study's purpose?

Sampling procedures seem to be reasonably well presented and there appears to be sufficient care in monitoring the attentiveness and legitimacy of individual respondents whether adult or adolescent. The sample's vendor appears to be especially careful about this. Kudos here.

The key analysis appears to be a comparison between responses to the new statements in the 16 experimental conditions to the old statements in the control condition on criteria such as

newness, perceived learning, and links between beliefs and outcome. Although this sounds like it makes sense, it's actually a weak criterion because the old statements have been rejected precisely because they are already well understood, and the new ones selected precisely because they are not so well known. I think I would've been more impressed if there was a no-exposure control and/or an inaccurate belief control to show that the new information is better believed than both the old information and the information that is a part of inaccurate claims.

The analysis plan for the knowledge learning and thinking about outcomes seems to make a fair amount of sense in that there is a comparison between items that are new and old but roughly matched on content between the treatment and control groups. This allows one to infer that the new version is effective on these three measures versus similar content in comparison to the old version. Where there is none whose content is comparable to that of the new than a random comparison is made between the new and the old which could easily overstate the effectiveness of the new. The authors recognize this, but it nevertheless does run the possibility of overstating the result.

5. *Are the stimuli used appropriate given the study's purpose?*

The FDA has chosen to study lesser-known health consequences in the new warning statements. The argument here is that more well-established and better known health consequences are already well-known and only need reinforcement not creation or conversion. The problem with teaching people something that is new is that "new information or claims" run the risk of being unpersuasive, and indeed raise skepticism about the new information given its novelty. We have run across this problem in several different contexts where what is new is less likely to be believed. This is a major issue here in Study 1 and in Study 2.

6. *Are the outcomes measured appropriate given the study's purpose?*

Public understanding which is clearly the goal here. It is being equated to self-reported learning and the newness of the information. So, if a warning statement is identified as new and is perceived as teaching someone something they didn't know before, then presumably this is a warning statement that is understood. This is an odd use of the word understanding which, in common parlance, is identified with comprehension and linkage with established knowledge. If I created an exam for students in my class and I asked them "is this information new to you and I asked them to report whether they learned something from the information" would I then conclude that they understood it? I think the answer is obviously no. Understanding is generally a concept that refers to the ability to use information successfully in one's life and to integrate the information with an established pattern of beliefs which is already accepted. So, I for one would find it difficult to equate these operational procedures with the ordinary concept of understanding or with the cognitive concept of understanding as used in the scientific literature.

It's clear that the revised warning statements outperformed the original TCA statements handily on criteria that really do not tap into understanding, acceptance, or knowledge other than as measured by self-reported learning. And as I argued above, these two measures do not tap into understanding in any sense of what the word understanding ordinarily means

conceptually, in ordinary discourse, or in scientific measures of comprehension. I am also not a fan of self-reported learning nor of novelty - that is awareness - as a criterion.

For this reviewer, the primary outcomes seem a lot less interesting than the secondary outcomes seem to be. My argument is that the knowledge, learning, and thinking about kinds of questions are transparent and in some ways don't really get at what their labels say they are getting at. For example, the abbreviated wording called new knowledge is really an awareness question.

Learning is really reported learning, not recall or understanding. The belief items are about the extent to which people agree or disagree with a claim which is actually a rewording of one of the warnings. Later questions assess whether people accept the rewording as factual or not and as causal or not. All the questions prior to this set of items are not really about acceptance and while this study is supposedly not about persuasion in reality it's crucially important for people to accept the warnings and not simply say they are new or say it leads them to say that they learned something.

7. *Are the study participants included appropriate given the study's purpose?*

One question about the weighting of the sample is how well the three age brackets reflect the distribution of smokers in the society. The selection process was one third under 18, one third 18 to 24, and one third 25 and older. The justification for this distribution is not clear but at a minimum should be compared to nationally representative samples.

In the sample demographics, I was surprised to see some significant asymmetries in male-female distribution by adolescent and young adult groups. Females significantly outnumbered males among adolescents and the opposite was true of the young adult sample. It's not clear why such sharp differences are present in the sub-samples or whether these differences might affect the results differentially for adolescents and young adults. I was also surprised to see in the sampling section that there was no discussion of how adolescents that were susceptible were defined. The definition is available later but should be there with the first introduction of the subgroup.

8. *Is the analytic approach appropriate given the design and purpose of the study?*

Table 3.4 offers pretty strong results for learning although clearly the results are primarily the result of adolescents and young adults in lesser so for older adults. Table 3.5 presents a strong evidence that new knowledge is enhanced for the new topical areas but no real advancement in terms of thinking about risks for these new topical domains except in a few cases (five to be precise).

Table 3.6 makes my point that the new statements are often seen as less believable even though they are newer and quotes informative and elevated an awareness of new information. But simply put, new is not necessarily acceptable; new is often less believable and that's borne out here. Similar findings are obtained with regard to facticity versus opinion.

9. *Are results presented consistent with the analytic approach?*

I appreciated the care with which the analytic plan was laid out and with the detailed attention to correcting for multiple comparisons and the presentation of both standard and corrected levels of statistical significance. Overall I think the analytic approach is not just solid but strong.

10. Are there any concerns with the results presented?

In Table 3.4 where the regression coefficients plus there are 95 percent confidence intervals are presented the authors use one of the kinds of presentations that drive this reviewer crazy. That is in presenting the confidence intervals they confuse dashes with minus signs whereas a simple modification could make it clear that something is a negative number versus something is a positive number by simply separating them with a comma; it's a trivial thing but it makes for clarity in presentation or at least the absence of confusion.

11. Are potential limitations of the study appropriately identified?

I have no doubt that these new warnings will work versus the non-existent warnings we now have. But that said, acquisition of knowledge and recall of the new warnings is not the same as accepting the information as true. So while the purpose of the warnings is not to persuade, it is nevertheless to have people learn in the sense of accepting information, to understand in the sense of being able to use the information and to integrate it into a complex of information that is a part of people's core understanding of the consequences of smoking combustible tobaccos. Simply being aware and saying that one has learned is not equivalent to having accepted the information. This is true in these data as well when the awareness levels (told me something new) are contrasted to the believability and facticity judgments.

It's very unfortunate that the allocation plan for the control group was not properly carried out. The reason of course is that the number of treatment conditions are so much greater than the control condition with equal allocation to all conditions without weighting. I had this problem in a study that I did a few years ago and rued the day when the treatment conditions were far out of proportion to the control condition. This could be a flaw in the study depending upon the kinds of analyses to be carried out in comparing some portions of the control group to various combinations of the treatment conditions.

There is no commentary in the sample description regarding the potential confound of gender with adolescent and young adult samples. The problem this obviously creates is that comparisons between these groups -- if any (none so far) -- will be confounded with gender. I suspect some weighting will happen as needed but the asymmetries are pretty strong in the sample.

12. Are the conclusions drawn from the study well supported by the data presented?

While it is certainly difficult and to some degree unfair to compare the revised warning statements to one another relative to the original TCA warning statements, at some point a decision has to be made as to which of these 15 should be prioritized if all are potentially eligible.

The following statement is crucial in the summary because it makes clear that believability

Quantitative Consumer Research on Cigarette Health Warnings: Study 2

I. GENERAL IMPRESSIONS

The comments that I have made in response to the 12 charge questions include various elements that would fall under “general impressions.” I draw them out here in a separate answer.

Both studies are very well done in terms of design and data analysis. The designs selected provide for an appropriate control groups which are the current standards for warnings (Surgeon General in Study 2) or immediate past selections (initial FDA warning labels in Study 1). The data analysis plan is strong and straight forward. It is careful on statistical treatment of the quality of data (e.g. continuous vs. rank order) and especially strong on correcting for experiment-wise error and power considerations. Both documents do a good job in communicating their procedures and the results except as noted regarding Study 1’s complex and difficult design that requires some supplementation with a clear and effective graphical description. If one accepts that the operational measures of learning and novelty are valid measures of their underlying constructs, then there is a clear picture that the new warning texts and warning labels are effective relative to their comparison.

However, I am concerned that the measures deployed – perceived novelty and awareness – are not convincing measures of the underlying constructs that the research is targeting. In Study 1, the researchers do employ a measure of believability and of facticity (opinion vs. fact) finding that the new labels are less believable. In Study 2 the believability measure is not present even though it was diagnostic in Study 1. Both the believability and facticity measures underscore the fact that the new warnings may not be accepted by the target audience. The authors are well aware of this and comment on it in both studies. But coupled with self-report measures of learning and novelty (awareness), the lower levels of acceptance of the labels reduce the overall impact of the results. The implicit rejoinder in the data to the argument that the results are not so convincing is that the warnings affect the acceptance of negative health consequences (i.e. beliefs) in the warning label conditions versus the SG warnings condition and do so over time. But as I note in my comments on Study 2, two concerns arise about these findings. The first is that asking beliefs at baseline before message exposure taints the message processing by focusing respondents’ attention to the messaging in ways that privilege the beliefs being targeted. Second, the beliefs are measured three times reinforcing the warning labels’ content. Third, it is not clear why all beliefs would be affected by a specific warning label as opposed to a more targeted set of outcomes wherein warning label X affects beliefs related to warning label X but not beliefs Y and Z.

In the end, these are both very carefully done studies that adhere closely to the data that has been gathered. This reviewer is raising interpretive considerations that essentially claim that the overall set of results are less convincing than they might be had the same constructs been operationalized differently and slight changes in the design in Study 2 been implemented.

II. RESPONSE TO CHARGE QUESTIONS

1. *Is the document logical and clear?*

Yes, the document is logical and clear. The study design is easy to understand especially in

contrast to Study 1 which is so much more difficult to figure out. The description of the design, the measures, the analysis techniques as well as the results are readily interpretable and readily comprehensible.

2. *Does the executive summary accurately reflect the content of the overall document?*

Yes, I thought the executive summary was very good and clearly highlighted the overall findings as well as giving a sense of how the data were gathered and the empirical procedures carried out. So, from the point of view of communication and presentation of findings I think the study too does a very good job. In spite of its length, it really does a nice job of presentation and communication.

3. *Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?*

The level of information provided is thorough in every segment of the work and so I have only minor quibbles about some additional pieces of information that might be provided. The stimulus materials that were presented were clear enough, but it was not clear how they were derived especially in the context of Study 1. Study 1 aimed to develop some new warning texts and the connection between Study 1 and Study 2 should have been obvious in some ways but was not.

Table 3-6 refers to completed surveys but needs to be more forthcoming by describing if this means completing all three sessions and what happens to those dropping out after the first session etc. Attrition is an issue and should be addressed in the sample completion section (somewhere for sure but this would be a good place).

4. *Is the methodology used appropriate to address the study's purpose?*

Since the design is longitudinal over three points in time there clearly will be attrition and indeed there was. Some careful discussion of the process of participant loss is called for and I make some suggestions about what to take a look at here.

Lightspeed's quality control mechanisms seem strong to me. As a researcher who uses online samples, I would be interested in using a panel with such quality control.

The items in the surveys are described in Table 4-1 and seem to imply that the set of 16 belief items were asked three times. This is fine except that asking these items at baseline distorts the way people process the information given in the labels cuing them into the content to be processed. In our message work we never ask the key outcome measures at baseline BEFORE the messages to be processed as we believe that that distorts how content is handled – priming, focusing, differential attention, etc.

Every design can be criticized for failing to do something. In the current design my greatest concern is that the belief items are asked three times. I'm not primarily concerned about test-retest sensitization which occurs in both the control and the treatment conditions but rather the interaction effect between the belief assessments and the follow-up warning labels. The problem is that the belief statements have content which is consistent with and primes the

respondents to focus on the warning label in a unique way. This creates the possibility -- indeed the likelihood-- that respondents are reacting to the content of the warning label in ways that they would not in the absence of pretest measures of negative health consequences. So, this can confound the results in ways that are different in the experimental context than they would be in the real world context where the beliefs are not primed systematically prior to exposure to the stimulus materials. This of course could have been designed out at substantial additional cost in terms of resources by having a post only condition to compare to the pre-post-conditions of the current design. But as I said every design can be criticized for failing to do something and jeopardizing internal validity.

5. *Are the stimuli used appropriate given the study's purpose?*

What led to the choices of the 16 given the results of the prior study? Why not stay with the original set? Why drop back to some of the previous warnings in the tested set? How do the texts developed from Study 1 play into the selections for Study 2? What did I miss?

Exposure to stimuli is not masked in any way. Respondents are exposed at time 1 and at time 2. Health beliefs are assessed at three points in time including prior to exposure at time 1. Asking all the beliefs together three separate times could have the effect of creating a clustering of beliefs such that any effects on one from a warning would transfer to the others even though that would be an unnatural result of the design and the set of items.

6. *Are the outcomes measured appropriate given the study's purpose?*

The believability criterion was not included in this study and that is problematic I think because these results undermined the legitimacy and utility of the warnings. Facticity was a problem with the new warnings in Study 1 and they are a problem here as well. The bottom line in this and in the prior study is that the new information is not as well accepted as the old information and so as a warning to smokers and potential smokers it will be less effective than a warning based on established claims. Of course, an established claim of negative health consequence is "old hat" and one cannot show improvements in learning as it is already overlearned. But new warnings will require support or will fall by the wayside in terms of their acceptability. The counterargument that this new warnings standard will be better than the existing warnings is without question going to be true but whether these new labels would be as effective as some established warnings that are already accepted is not tested and is a legitimate counter hypothesis. Another counterargument from the data of this and the prior study is that beliefs are affected by the new labels and so are actually pretty effective? But the comparison establishing this claim is a weak set of existing SG warnings and the beliefs shown to be affected are ones that are variations on the wording in the new warnings already (and repeated multiple times). So, the counterargument is a weak one I think. Testing should have included some false beliefs or beliefs not a part of the warning set to show that the effects of the new warning labels are on the targeted beliefs and not a general halo on any and all smoking related beliefs.

7. *Are the study participants included appropriate given the study's purpose?*

Sample selection. The adolescent group included smokers and susceptibles which makes sense given that the transition to smoking occurs early and not later in life. But why only use smokers and not-susceptibles in young adults where the transition still is occurring. Makes

sense not to include susceptibles among adults but what is the justification for non-smokers in adults (former smokers yes but never smokers)? What will this population tell us about the effectiveness of warnings other than the political acceptability of the warnings?

The sample description in Table 2.2 says that adult non-smokers currently “do not smoke at all” but does this mean that they could be former smokers? Same for young adults. This is definitely clarified later on in the document, but it would be reasonable to make sure that it’s clear at the outset.

In the sample’s demographics, the adult sample has a significant asymmetry in age distribution with 35-55 underrepresented.

8. *Is the analytic approach appropriate given the design and purpose of the study?*

What is not clear in the analysis plan of beliefs is whether any treatment vs. control exposure is expected to affect any and all beliefs tested. Obviously, the warning label in a condition only refers to particular beliefs and not all beliefs. So why should any other belief not referenced in the warning be affected by the warning? The partial answer to this is spillover through cognitive activation but without some idea of which beliefs are correlated with which other beliefs, there is no way of anticipating what is and is not affected by spreading activation. Not clear what exactly is being tested in the belief analysis plan e.g., all beliefs regardless of condition or all beliefs but as a function of condition. Why would the impact of one specific warning affect with a narrow focus affect all the other non-congruent beliefs?

9. *Are results presented consistent with the analytic approach?*

The effects on beliefs from session 1 to session 3 are of course even more difficult to produce statistically significant effects but there is indeed some evidence of the remaining effect on beliefs even as long as two weeks out. Obviously, there are always competing explanations having to do with attrition rates and test sensitization among other things. But these are potentially strong findings from such a small and unfamiliar warning dose.

10. *Are there any concerns with the results presented?*

The strong effects obtained on judgments that the information presented in the new warnings is new would have been more persuasive to me if for example some false information was included but was not a part of the warnings or information that was not a part of the warnings was included. In order to see that the “thinking about risk” outcome is to some degree not completely dependent on the “new information” judgment which it follows in the questionnaire, some measure of association between the new information measure in the thinking about risks measure should be provided. What is also clear is that the perceived factualness of the warning is less accepted when the information is seen as new. So just as in the Study 1 people identify information that they hadn’t seen before as new but also less likely to be a fact rather than opinion and therefore less likely to be accepted which is a part of the learning process. What happened to believability?

The fact that the recall item at session 3 was better recalled than the specific textual warning from the Surgeon General’s warnings is not at all surprising. The respondents were exposed

to their specific warning within condition four times – session 1 and session 2 one on a cigarette pack and the other on an advertisement-- so to find that the recall of the text that they saw and read after four exposures versus no exposures is minimal evidence of the recall ability of the materials. Also, this is a recognition test not a recall test. Recognition tests are a lot easier than recall tests would be. A third factor of course is the presence of a visual image supporting the text in a way that is presumably concordant with the text that is being recalled. So, several factors argue in favor of successful recall that is recognition of the message to which they were exposed.

11. Are potential limitations of the study appropriately identified?

In discussing the demographics of the sample at times one through three I think it would be reasonable to compare the characteristics of the sample to more representative samples that have been gathered by other sources simply as a way of saying that the sample is close or far from established samples of smokers and susceptibles among young people.

The attrition from time one to time three is substantial as would be expected but it is incumbent upon the researchers to discuss a little bit about who it is that remains in terms of things like a priori beliefs and how those change over time with the attenuated samples of sessions two and three. For example, adolescents who are susceptible to smoking seem to drop off from session 1 to session 3 and the percentage of adult non-smokers seems to increase from session 1 session 3. So, by session 3 the susceptible adolescents are down, and the adult non-smokers is up. The attrition by condition does not look appreciable to me on any criteria other than the couple of demographic differences that I noted above. But one of the things that makes a lot of sense to me is to report attrition as a function of session 1 beliefs. So, for example if there is evidence to show that those who are more accepting of the negative health consequences remain in the sample than they are going to be more likely to be attuned to the messages that the warnings carry being more engaged attentive and accepting. If that is the case then the very favorable outcomes observed are overstated.

12. Are the conclusions drawn from the study well supported by the data presented?

Let's focus on the belief changes from session 1 to session 2 in contrast to the changes in the control condition. What's clear from the results of Table 3.5 is that acceptance of the beliefs that smoking (for example) causes bladder cancer is elevated in the treatment conditions in contrast to the control conditions. But why would this be? First of all, the treatment conditions mentioned bladder cancer at both session 1 and session 2 but only in the warning condition that's about bladder cancer. In the other 15 conditions there is no discussion of bladder cancer so why would the effect here on accepting bladder cancer as being caused by smoking be so distinctive across conditions when in fact only one condition mentions bladder cancer? Should it not be the case that the treatment condition mentioning bladder cancer should carry the weight of the impact and the other conditions mentioning other health consequences not be affected or at least affected to a much lesser degree. If that's not the case then there is something else going on where in the warning about bladder cancer is dragging the other effects along with it on other health consequences and so the content of the warning is less consequential to belief change than one would expect. Clearly, if some form of spreading activation among potential beliefs is the basis for these effects then the particular warnings don't matter as much to the effects on beliefs or there is some process

other than the warnings that is producing these effects. I would be more convinced that it's the content of the warnings if indeed what happened was that the bladder cancer warning as the greatest effect on the change from session 1 to session 2 than the other conditions do; alternatively that the bladder cancer condition affects the bladder cancer belief but not the other health consequences or does so to a lesser degree. This is easy to check and would make a stronger case that it is the warnings that produce the change rather than some other process obviously related to the warnings but not necessarily consistent with the content of the warnings.

In the conclusion section, the authors argue that the skepticism attached to the graphic health warnings because they are new might disappear over time as exposure is increased. That may be true, but I think the reality is that a persuasion oriented campaign consistent with the claims that are being made here providing more elaborated evidence, information about credibility, and even testimonials to support the claims may be necessary. Otherwise, the warnings may be seen as new and unfamiliar and remain in the domain of opinion rather than fact, and less believable and accepted than would be desirable.

III. Specific Observations on *Experimental Study of Cigarette Warnings: Study 2 Report*

None provided.

External Letter Peer Review of Quantitative Consumer Research on Cigarette Health Warnings Required by the Family Smoking Prevention and Tobacco Control Act

Reviewer #4

Quantitative Consumer Research on Cigarette Health Warnings: Study 1**I. GENERAL IMPRESSIONS**

This extremely clearly written report documents an online experiment that was conducted to evaluate the efficacy of new smoking warning statements (versus older statements) for improving knowledge of the health risks of smoking. The information is accurately reported, and a very good amount of detail is provided to support transparency in reporting the methods and results. The methods mostly follow logically from the stated goals of the research, and the complex experimental design is suited to the task at hand. The report carefully describes power analysis, rationale for taking specific analytic approaches, and drawing conclusions based on the results (essentially a restatement of those results). The report was lacking in several areas and would benefit from: providing more detail and rationale on the need for warning labels and conceptually, the link between warnings and behavior or behavior change; being clearer on the origin of the assessments and their validity; and providing more information on vendor and sampling issues. Presentation of the results would be improved with more frequent summaries and inclusion of graphic presentation of key results (e.g., for primary outcomes). It is unfortunate that the study was not powered to detect differences in different subgroups or by age; warning statements are not received the same by some segments of the population and to the extent that the warning labels need to cover a wide population base, it would be important to know what worked and for whom. The key limitation, for this reviewer, is in the sampling frame and sampling design. The report acknowledges the limitations with employing a convenience sample (albeit one that is very diverse), but the robustness of the conclusions depends, in part, on the extent to which they generalize to the population as a whole; and this is just not as easily possible with a convenience sample. Additional information on vendor choice, features of their panel, and decisions to not utilize weighting or some other sampling scheme would help to contextualize the results.

II. RESPONSE TO CHARGE QUESTIONS***1. Is the document logical and clear?***

The report of Study 1 is very well-written, and clearly described. Some additional details on study rationale, features of the design, and analyses would have improved the depth and quality of the report, however (these are described in subsequent sections of these review comments).

2. Does the executive summary accurately reflect the content of the overall document?

Yes. The executive summary is quite detailed and provides sufficient information that accurately reflects the report as a whole.

3. Is sufficient information provided about the study design, stimuli, sample, methods,

analysis, and results?

The study design, analysis, and results are described and are sufficiently detailed. Aspects of the sample, stimuli, and methods are less detailed. Some of the information needed includes (these are described in more detail in subsequent sections of this review): more conceptual background and motivation for studying warning labels; the value of warning labels; and need for new labels (beyond statutory requirements); formative work that went into development of the new warning labels; decision rules for segmenting the sample (e.g., using days smoked in past month; excluding committed never-smokers) and for choice of vendor (given the important limitations inherent in the non-representative sample); and validity of key measures used.

The study is framed in a very practical manner, i.e., TCA developed text warnings, this study is designed to improve on those warnings by providing additional detail. As important as this feature is for ease of communication, it is lacking in conceptual and empirical motivation; the result is that the rationale for the study comes across as thin. An enhanced background section (it does not necessarily require pages of dense theory and analysis of previous research) – one that reviews relevant literature on health warnings, particularly the efficacy of such warnings (and need to change them periodically/frequently); importance of warnings for improving knowledge; importance of knowledge to motivating behavior change, etc. – would really improve the scholarship of the report and motivation for the study. Such information would also help to motivate the study hypotheses (which appear later).

4. Is the methodology used appropriate to address the study's purpose?

The study design is complex but given some of the built-in limitations of internet survey panels that challenge internal validity of experiments, the design was well-appropriated for purposes of this study. In other words, the randomization scheme and viewing allocations and random ordering of study stimuli in the control and experimental conditions (as well as the data security checks) improved internal validity and helped to overcome the limitations with internet panels.

5. Are the stimuli used appropriate given the study's purpose?

The study has a very focused purpose, and to the extent that the experimental stimuli reflect warnings that will actually be used and (potentially) implemented, these stimuli are appropriate to the task at hand. It makes logical sense to compare these newer warnings to the TCA warnings provided to evaluate whether those newer warnings improve knowledge and retention over and above the prior ones.

Information on the development of the revised statements is lacking but is needed to evaluate their conceptual adequacy and literacy level, among other issues. More description of the formative work that drove the development of these new warnings is needed.

6. Are the outcomes measured appropriate given the study's purpose?

The measures provided were seemingly drawn from extant assessments (as generally asserted in Section 2.3) but there are no links between these references and the actual measures used.

Providing these links (as well as the rationale for such choices) is important for determining validity of the items and whether they were used appropriately in this study. For example, does improving knowledge lead to behavior changes? Providing a link to actual behavior is, of course, outside of the scope of the goals of this project. However, to the extent that knowledge is a proxy (or alternatively, mediates behavior change from warning exposure to behavior), it is important to consider and discuss.

The single item assessments (e.g., informativeness; believability; fact vs. opinion) would necessarily be lacking in reliability (versus a longer scale). Disposition of these items and any validity evidence would help to underscore their appropriateness.

Other measures could have been employed, even in the context of this internet panel. For example, memory of risks, true-false items, etc. The stated measures are likely adequate (especially if evidence can be provided on their psychometric soundness and use in previous studies) but other assessments could have been used to further meet the study goals.

The decision to parse outcomes into primary and secondary was not clearly stated; or put another way, the reasons that some outcomes are considered primary versus secondary were not clearly stated. For example, given the stated goals of the study, “number of health conditions” seems more like a primary versus secondary outcome. Other secondary outcomes (believability, factuality, informativeness) seems tangential to the primary goal of improving “understanding of the risks”. Additional rationale is needed on these points.

7. *Are the study participants included appropriate given the study’s purpose?*

The sample employed for this study is one of convenience, which brings with it a host of potential biases and limits to generalizability versus employing a representative sample. The report describes limitations with this approach to sampling versus other options (Section 4.2) and describes cost as being a core decision-making factor in choosing convenience instead of representative sampling. There are differences in vendors in terms of sample quality – so one wonders about the reasons that Lightspeed was used versus some other vendor. Were there substantive reasons to select this vendor over others beyond cost and some of the data security features described (Section 2.2)?

It would be useful to know what potential participants were told during recruitment. Were they given full details about the study or were smoking, warning labels, FDA, etc. mentioned to potential study participants during recruitment? Was a cover story used? This point is important because study framing could have an impact on those that agree and refuse; and also help to understand more about potential sampling biases (e.g., were those that agreed more likely to do so because they already held knowledge about smoking or needed knowledge?).

One also wonders whether any methods (e.g., weighting) were or could have been employed to increase representativeness and generalizability.

That said, the sample is diverse, so this offsets lack of representativeness and potential biases related to sampling (and match to population characteristics) somewhat. Still, lack of representativeness is probably the most serious limitation with this research.

Quantitative Consumer Research on Cigarette Health Warnings: Study 2

I. GENERAL IMPRESSIONS

This very clearly written report describes a study that compared new graphic health warning labels for cigarettes to those that are already in circulation in the U.S. on various measures of health education. The methods and results are very clearly described in a way that lends the study to great transparency in its approach. The conclusions and reporting of results are accurate. The randomized design is well described and appropriately conducted to ensure internal validity (especially important with an internet panel experiment). Concerns that detracted from the overall quality of the study and report include the following. First, the study needs greater levels of conceptual and empirical motivation – even as a practical matter, it is important to understand the reasons that graphic health warnings are important and why improving knowledge of the health risks of smoking could lead to changes in behavior. Second, the measures, while clearly described, are not linked to any specific study or research program, making it difficult to evaluate their utility and validity (beyond face validity). Third, the sampling scheme is a major limitation as the convenience sample limits generalizability of the study findings and the attrition rate over the three study sessions (over almost three weeks) is very problematic. Finally, the stimuli are necessarily artificial in experiments such as this one, but some design decisions (use of a blue package; model in the mocked-up design) could limit the overall generalizability of the study findings (i.e., for non-blue packages; and ads that do not feature a musical theme or male model).

II. RESPONSE TO CHARGE QUESTIONS

1. *Is the document logical and clear?*

The document is exceedingly clear in its purpose and in its style of communicating. It follows a logical course (with some redundancy between sections adding to the readability of this dense, comprehensive report). Level of detail, particularly as it relates to the analytic plan and results, are crystal clear – supporting transparency of this endeavor.

2. *Does the executive summary accurately reflect the content of the overall document?*

The detailed executive summary accurately reflects the content of the full report.

3. *Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?*

The design and analyses all are very clearly described. The sample is clearly described, but some additional details would help to overcome structural limitations with the sampling frame (i.e., a convenience sample is used vs. a population-representative sample). The specific procedures that went into development of the new text warnings and images were not described in detail; more information on formative work would improve the presentation of the report. Additional information that links measures to their origin in the literature and provides some evidence of validity is needed. Results would be improved by providing graphical representation of the results. Overall, the study requires more conceptual motivation other than to build upon the practical motivation of testing labels as required by

statute. All of these issues are described in detail below.

4. *Is the methodology used appropriate to address the study's purpose?*

The randomized design is straightforward and enables a mostly internally valid test of the study hypotheses (about the primacy of graphic health warnings over text-only warnings).

5. *Are the stimuli used appropriate given the study's purpose?*

From a practical perspective within the context of the study goals, the stimuli are appropriate. The present study is linked to the former study (OMB# 0910-0848) but it does not appear as though the results of that former study were used to inform the stimuli choice in the present study. In other words, the most effective stimuli (text warnings) from the former study were not used in the current study; all of the text warnings were used. This is not a limitation or criticism per se, but it does suggest that the relationship between studies needs to be more clearly explained.

In addition, the development process for the graphic images needs to be more clearly described. What was the formative testing process that went into designing and then finally choosing those images? What was the charge given to the designer that developed those images? Were images keyed to the specific text warnings? Were issues regarding diversity of the model facsimiles part of the development (e.g., regarding age, gender, apparent race or ethnicity)?

There is an element of artificiality in this kind of experimental study, of course, but the decision to use a cigarette named “brand” versus a name brand was not clear. There also could be concerns about use of a blue box for this cigarette brand stimuli as different colors communicate different features of the product. For example, some older research by the tobacco industry suggests that blue could convey information that the product is “safer” or more geared toward males (again the research is old). While there were not differences between conditions in the package color, the findings may be different with differently colored or logoed packages (or packages for real brands). Some consideration of this issue is warranted.

Decision processes and development of the print ad is also needed (e.g., why music? why a male model?).

6. *Are the outcomes measured appropriate given the study's purpose?*

The outcomes are geared toward the goal of the study, but more information on the origin of the measures and validity needs to be considered. There are general references provided in the aggregate for the measures but linking specific measures to a specific citation is important for determining the appropriateness of that measure.

7. *Are the study participants included appropriate given the study's purpose?*

The sampling frame is a significant weakness for this research. As noted in the report, convenience samples suffer from a host of structural biases that cannot be easily overcome.

(unless there is some sort of weighting procedure considered; but was not apparently done for this study). Moreover, the loss to follow-up over the short time frame of the study creates additional problems. The report advances some discussion points to minimize the attrition problems; but these are significant problems, nonetheless. Some analysis of drop-outs versus completers, sample representativeness, etc. would help to offset some of these concerns about the sample.

Some information on recruitment methods are needed. For example, what were participants told about the research? Were participants kept blind to the study purpose up to a point? These issues are important to consider because each has implications for contextualizing the limitations to generalizability of the sample.

8. *Is the analytic approach appropriate given the design and purpose of the study?*

The analytic plans represent a solid test of the study hypotheses. Correction for multiple tests is considered, and the results are presented and discussed plainly. Key results could also be presented via figures to improve communication of this densely packed set of results.

9. *Are results presented consistent with the analytic approach?*

The analytic approach was very clearly described, and the execution and presentation follow from this approach.

10. *Are there any concerns with the results presented?*

No concerns beyond the fact that the study was not powered to detect subgroup differences. While the report is very clear about this decision, it is unfortunate because to understand what works and for whom (and potentially why) are important questions given widespread distribution of the warning labels.

11. *Are potential limitations of the study appropriately identified?*

Major study limitations are described in the conclusions. Some of these limitations are quite serious (e.g., sampling) and the report does a reasonable job of explaining and describing implications of these limits.

12. *Are the conclusions drawn from the study well supported by the data presented?*

The conclusions are generally non-numeric restatements of the findings and are accurately stated. However, some of the non-significant results for some of the warnings are discussed and explained; these sorts of explanations are largely speculative (if informed) and also do not match with other aspects of the report (where non-significant results are left unexplained). It may be best (most consistent) to not offer such explanations.

III. Specific Observations on *Experimental Study of Cigarette Warnings: Study 2 Report*

None provided.

6. *Are the outcomes measured appropriate given the study's purpose?*

Overall, the outcomes are appropriate. However, I have a concern about the collinearity of two of the secondary outcomes. I believe that they may be highly correlated and could either be combined or one could be dropped if they are highly correlated. I did not have access to a correlation table, but that is my suspicion.

7. *Are the study participants included appropriate given the study's purpose?*

Yes, I think it was very appropriate to assess youth susceptible to smoking, youth smokers, and adult smokers. In other words, I agree with not including youth who are non-susceptible nonsmokers and adults who are nonsmokers.

8. *Is the analytic approach appropriate given the design and purpose of the study?*

Although I am not a statistician, I believe that the statistics were appropriate. The authors conducted a large number of statistical tests and performed analyses to avoid making Type 1 errors. I am aware of Bonferroni corrections, but not the FDR method; however, it seemed appropriate from the description.

9. *Are results presented consistent with the analytic approach?*

Yes.

10. *Are there any concerns with the results presented?*

No.

11. *Are potential limitations of the study appropriately identified?*

The study sample as a convenience sample is a limitation but that was clearly identified. The study also relied totally on self-report and that could have been highlighted a little more clearly.

12. *Are the conclusions drawn from the study well supported by the data presented?*

Yes, the conclusions generally flow well from the data presented. However, the conclusions are very conservative. The report does not really 'take a stand' on which warnings might be the best according to the data. I prefer findings that are a little more prescriptive based on the evidence and that synthesizes rather than merely regurgitates the findings. That said, I do not think the report goes well beyond the data.

III. Specific Observations on *Experimental Study on Warning Statements for Cigarette Graphic Health Warnings: Study 1 Report*

Yes. The methods section was thorough and detailed. I do think that a quick summary of the planned vs. actual assignment to condition error would be useful in the results section. The explanation of the effects of the error were clear.

4. *Is the methodology used appropriate to address the study's purpose?*

Yes. One drawback to the method is that some of the labels had multiple health beliefs and others only one. Because these are treated differently in analysis (some linear, some ordinal), it makes comparisons about label features more difficult.

5. *Are the stimuli used appropriate given the study's purpose?*

Yes. The stimuli are appropriate. Using the statutory and revised statements allow for a direct comparison and allows for conclusions about which ones are better suited to increase knowledge about the harms of tobacco.

6. *Are the outcomes measured appropriate given the study's purpose?*

Yes. It would be helpful for there to be some theoretical expansion on some of the constructs measured. The rationale for the new knowledge, learning and health beliefs is clear, but I would have liked to have seen some more rationale for including thinking about the risks, believability, and factuality. Why are these particular measures important to the purpose?

7. *Are the study participants included appropriate given the study's purpose?*

Yes. You may wish to include a brief rationale for not including adult non-smokers, I believe this is a strength of the study, as these statements are designed to appear on packs and therefore the intended audience is smokers.

8. *Is the analytic approach appropriate given the design and purpose of the study?*

Yes. While the reasoning behind using ordinal and linear regression based on whether the items could be scaled is scientifically sound, it does make the interpretation of the results more difficult. The sample seems large enough to consider dichotomizing the responses as agree/disagree. If the pattern of findings holds, it would be easier to interpret, and would help the reader draw conclusions.

9. *Are results presented consistent with the analytic approach?*

Yes. The results section provides a great deal of information on the individual statements and warnings, and it is consistent with the proposed analytic strategy. However, it would benefit from a few more summary figures or tables that help highlight the results of the slightly bigger picture question being asked. What do these results tell us about the statutory labels vs. the revised labels as a whole?

10. *Are there any concerns with the results presented?*

No. If there is a way to summarize the results of Tables 3-6 and 3-7 in an additional table,

Yes. You could consider giving some details about the development of the photorealistic images and how they were determined as concordant with the messages.

6. *Are the outcomes measured appropriate given the study's purpose?*

Yes. It would be helpful for there to be some theoretical expansion on some of the constructs measured. The rationale for the new knowledge, learning and health beliefs is clear, but I would have liked to have seen some more rationale for including thinking about the risks, believability, and factuality.

7. *Are the study participants included appropriate given the study's purpose?*

Yes. I would clarify whether the non-smoker category includes never smokers and former smokers, or only former smokers. Since you are including non-smokers in this study (as opposed to Study 1), you might provide a brief rationale for the inclusion on non-smokers as a group in this study and any hypothesized differences.

8. *Is the analytic approach appropriate given the design and purpose of the study?*

Yes. I'm slightly concerned about the number of times the participants were exposed to the health beliefs assessment and whether there could be an interaction with the condition. All participants were exposed to a large list of health effects of cigarettes that are, by design, novel. Then only those in the treatment condition are being exposed to the warnings with the new/novel information on it. It concerns me slightly, that there might be a priming effect of the health belief assessment for those participants in the treatment condition.

9. *Are results presented consistent with the analytic approach?*

In the presentation of results, rather than order the tables by condition, which has no inherent meaning to the reader, consider leaving the control at the top and then ordering the remaining rows by the highest (new information, or highest mean). It allows the reader to see the labels from most effective to least effective on some of the main outcomes.

10. *Are there any concerns with the results presented?*

Potentially. In addition to differential attrition by condition, I would like to make sure there wasn't differential attrition among key demographic groups between time points, it wouldn't impact the treatment/control findings, but it might hamper the generalizability of the findings if particular demographic groups were more likely to drop out. Looking at Table 3-1 it appears that low income individuals were dropping out more than higher income individuals, and that higher education participants were retained better than lower education. I would be curious to know if these demographics dropped out at different rates per condition.

11. *Are potential limitations of the study appropriately identified?*

Mostly. I was interested in the possible impact of an interaction between testing and condition. All participants saw a list of health beliefs (many of which are considered novel), and then those in the treatment group saw a warning that potentially reinforces one of those